

PROJET DE FORMATION



DataScientest.com



Présenté par Diego Guzman & Danyl Delaisser



OBJECTIF DE CITY WALKER

CRÉER UNE APPLICATION OPTIMISANT UN ITINÉRAIRE TOURISTIQUE



Automatisée



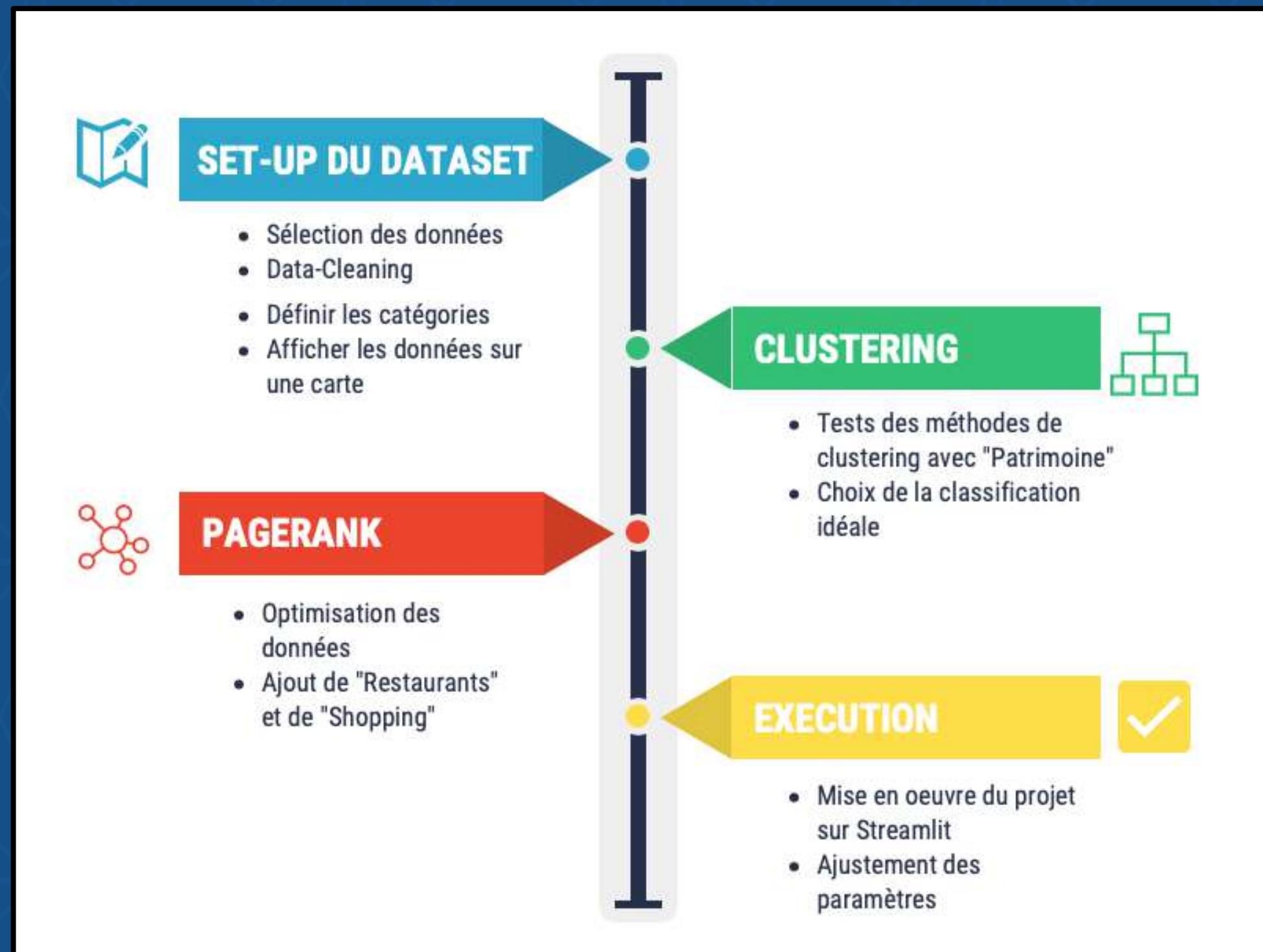
Intuitive



Personnalisable



ÉTAPES DU PROJET



CRÉATION DU DATASET



Partir de 0...

Catégories :

- Patrimoine
- Restaurants
- Commerces

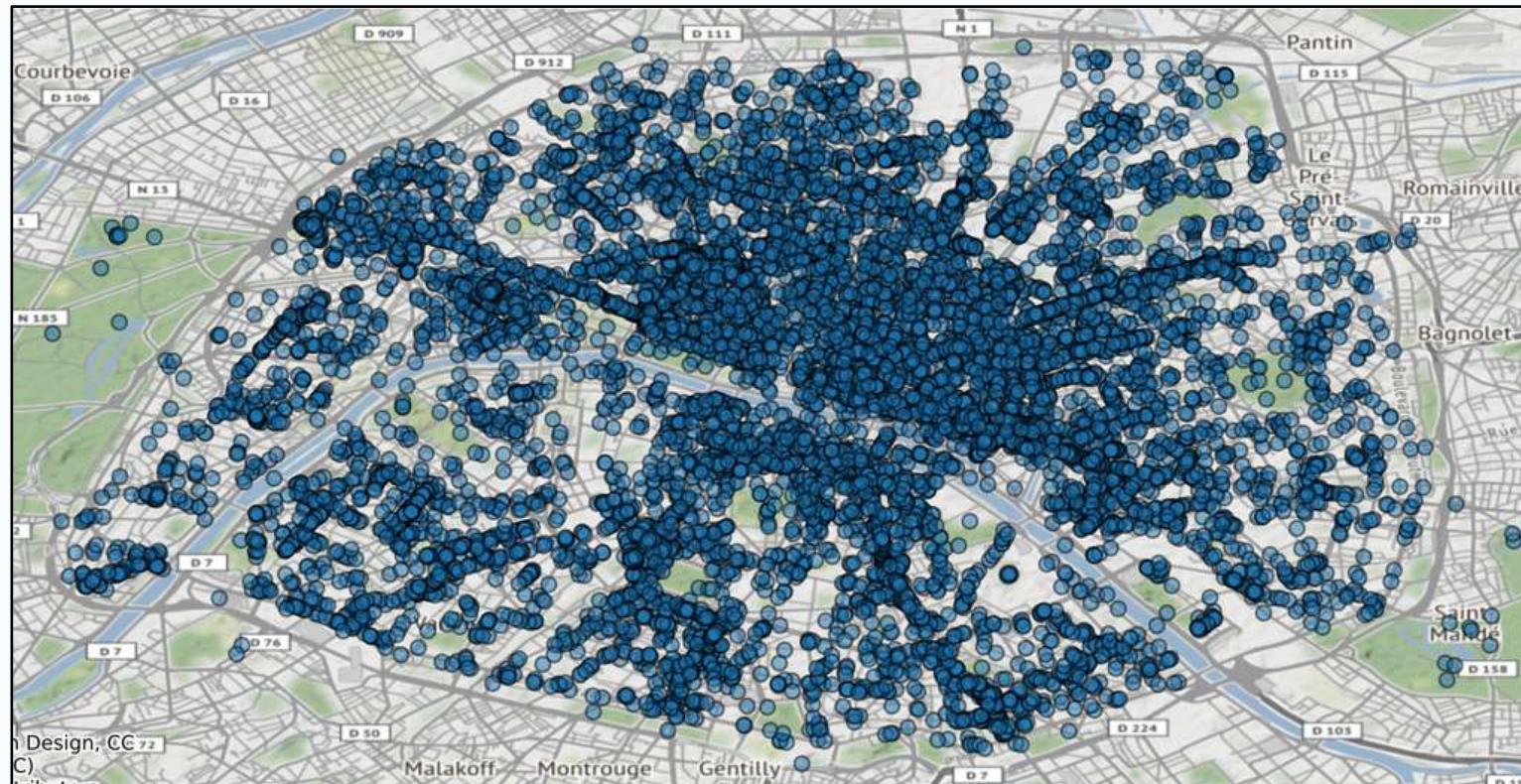
35 000 lignes



10 000 lignes



VISUALISATION



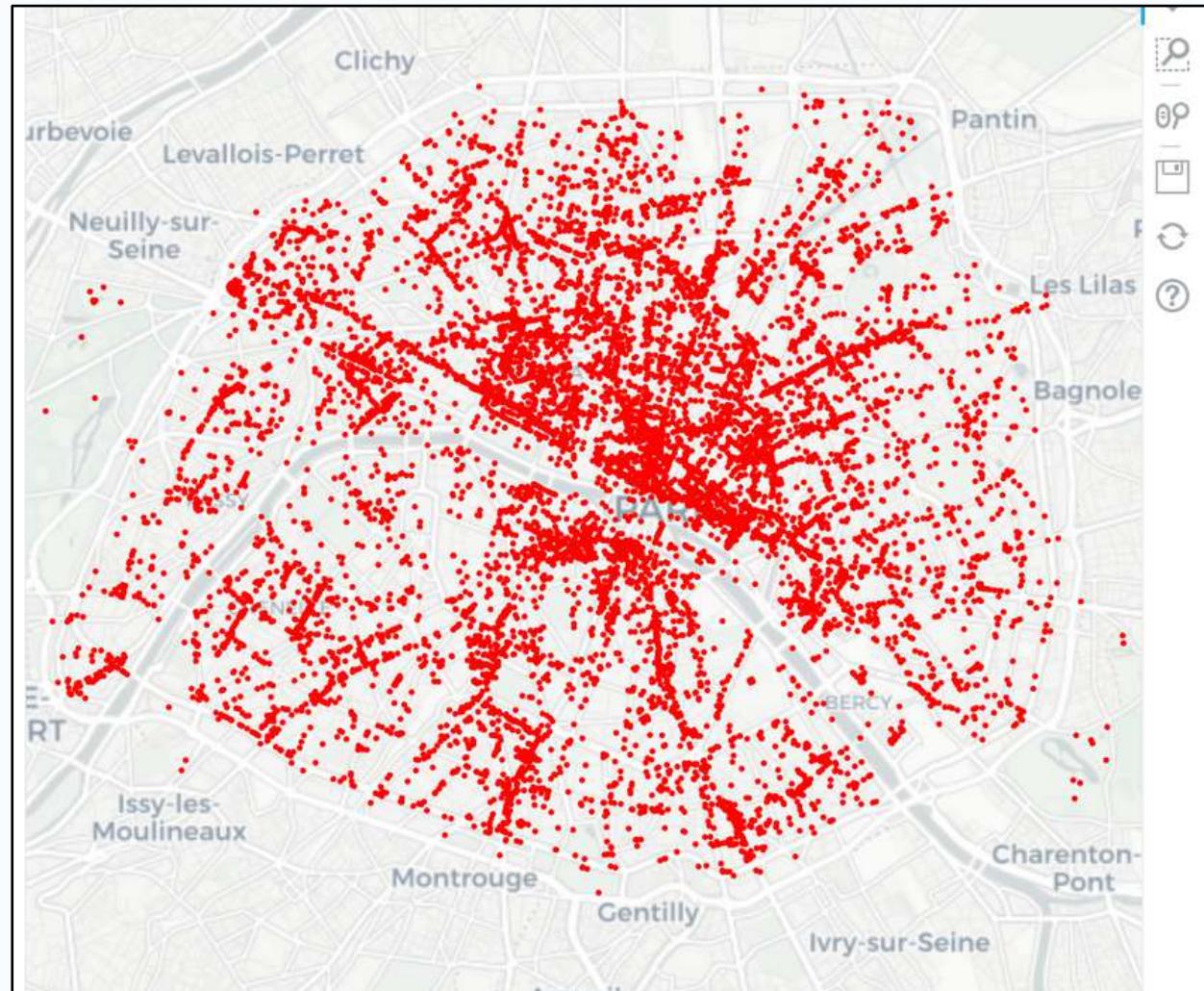
GEOPANDAS

+

- Outil accessible
- Prise en main intuitive

-

- Cartes fixes
- Difficile de trouver des ressources en ligne



BOKEH

+

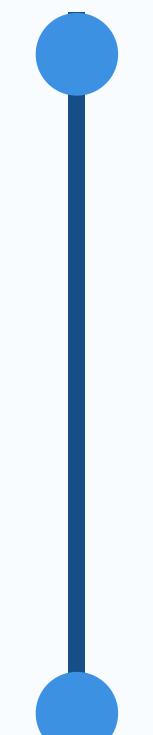
- Interactif et dynamique
- Possibilité d'onglets

-

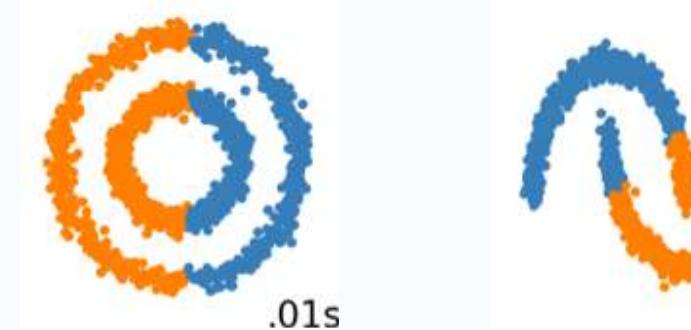
- Compliqué à coder (Arguments individuels)
- Difficile de trouver des ressources en ligne

METHODES DE CLUSTERING

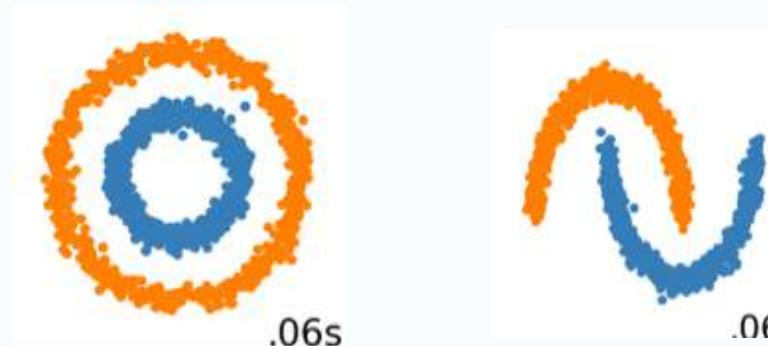
CATÉGORIE
CHOISIE
=
PATRIMOINE



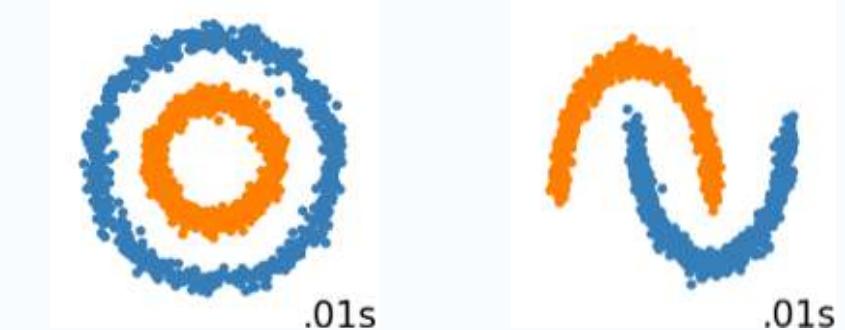
K-MEANS



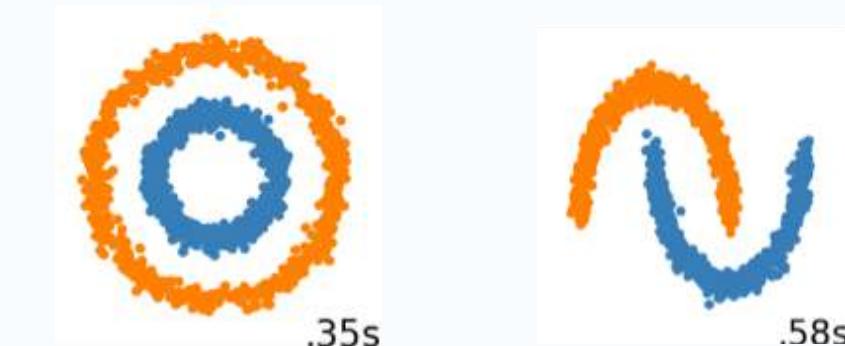
AGGLOMERATIVE
CLUSTERING



DBSCAN

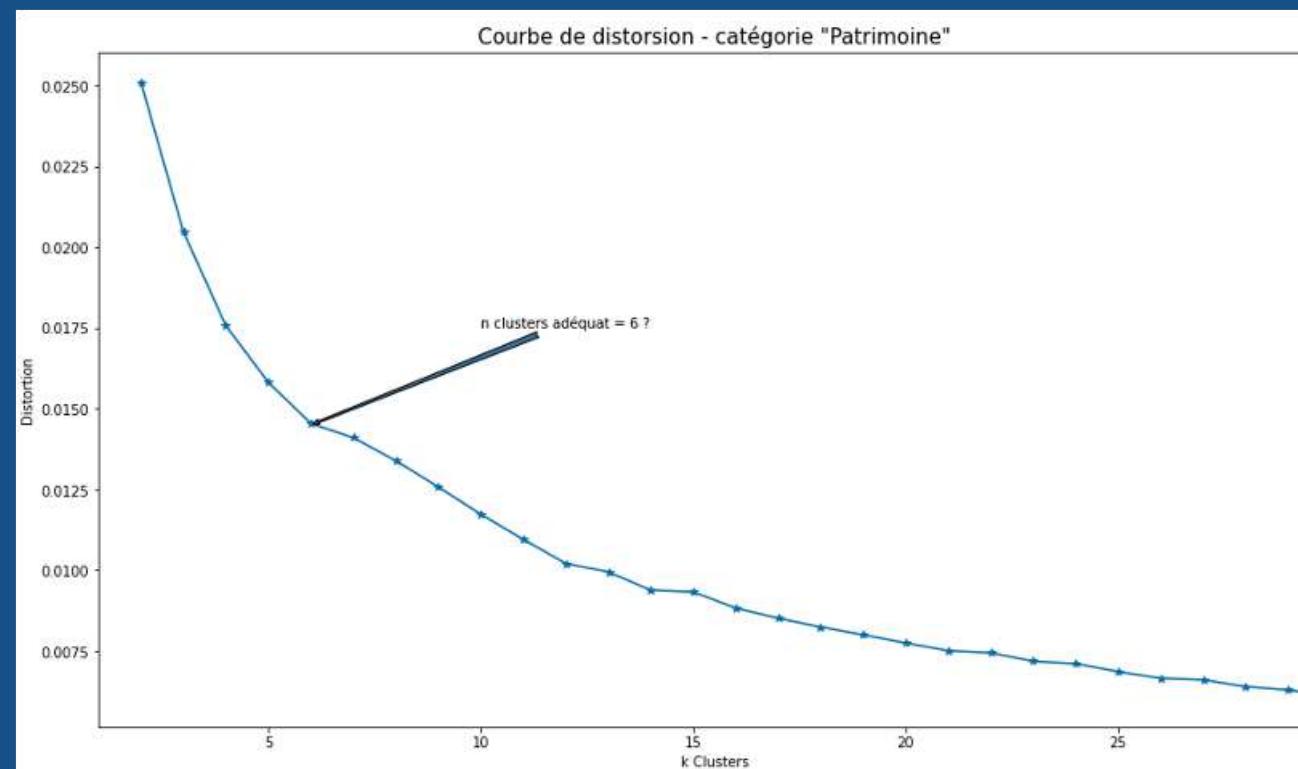
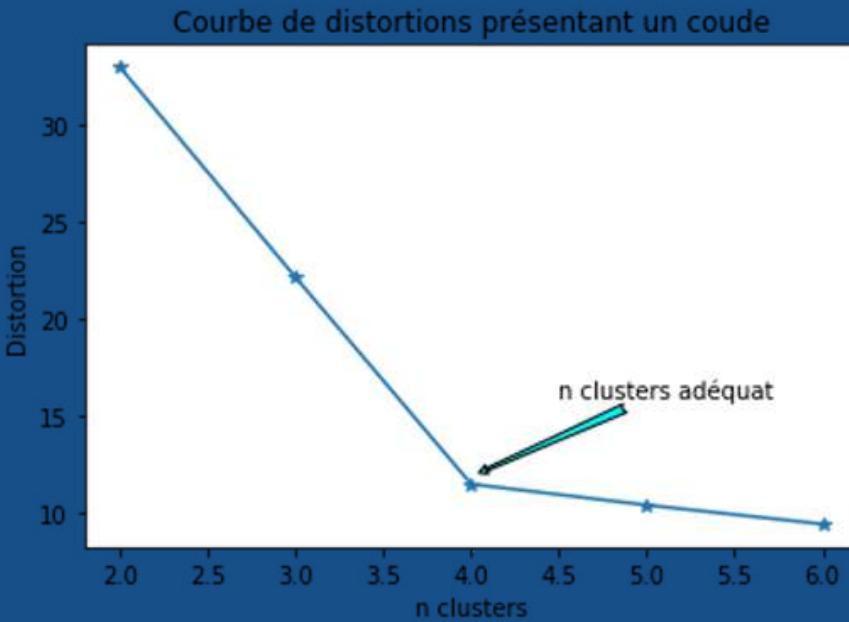


SPECTRAL
CLUSTERING



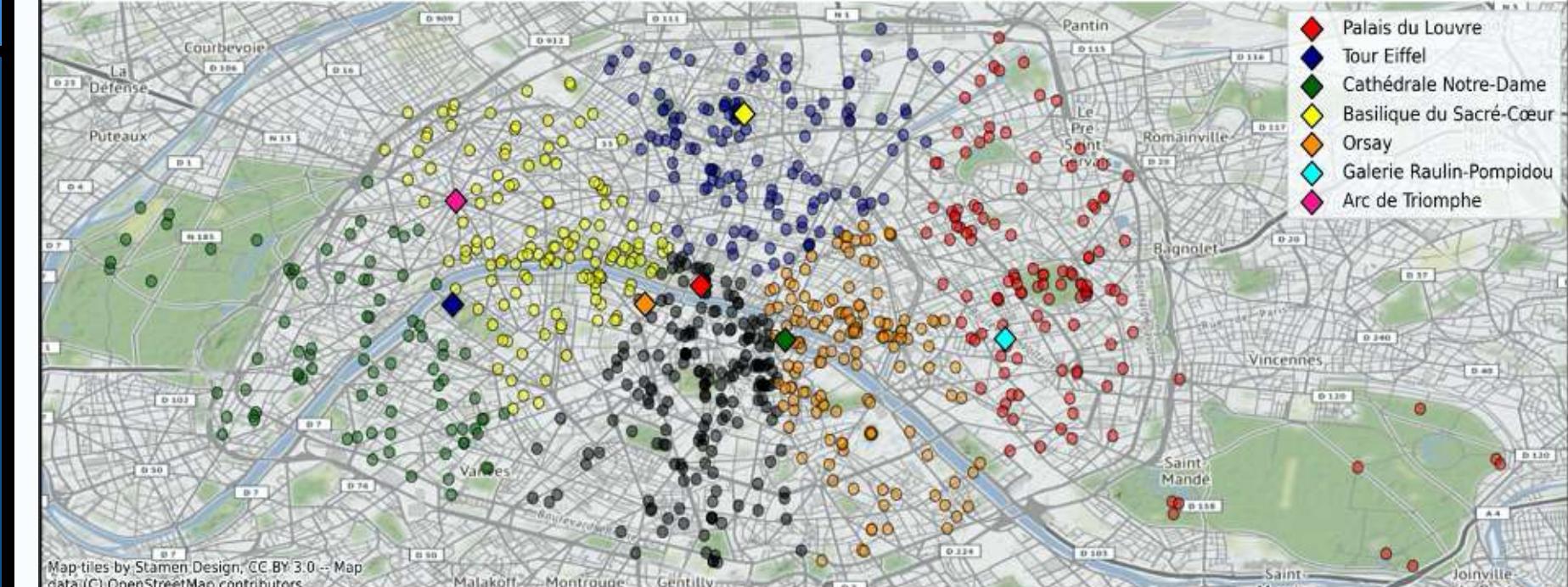
K-MEANS CLUSTERING EN 2 ÉTAPES

1 - MÉTHODE DU COUDE



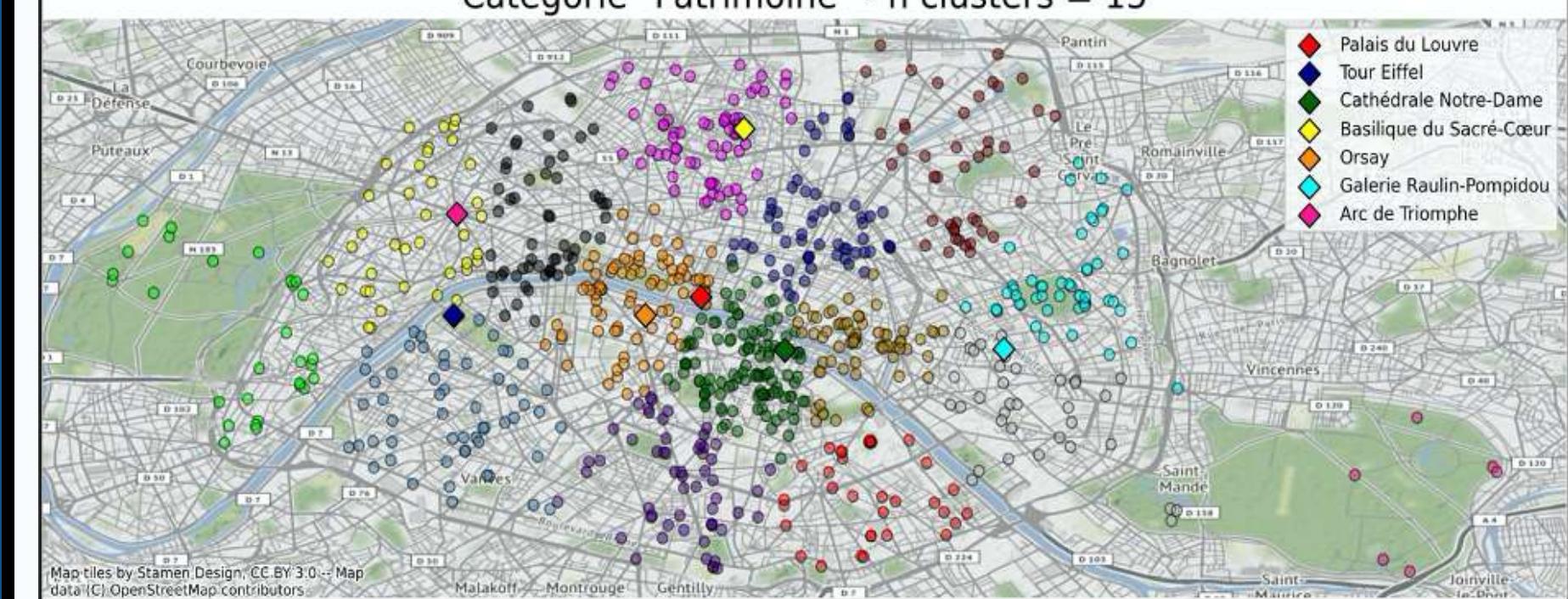
2 -CLUSTERING ET VISUALISATION

Catégorie "Patrimoine" - n clusters=6



Augmentation des n_clusters

Catégorie "Patrimoine" - n clusters = 15



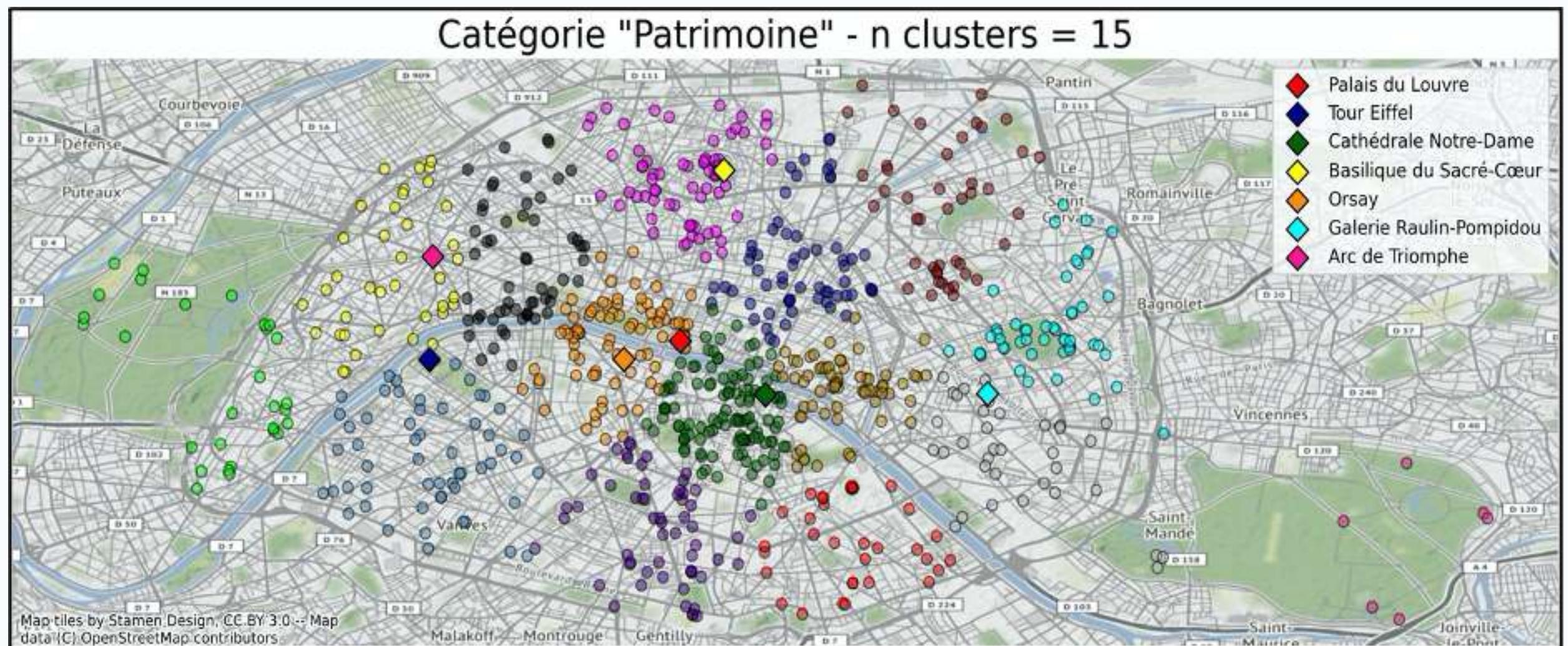
K-MEANS CLUSTERING EN 2 ÉTAPES



2 - CLUSTERING ET VISUALISATION

- Répartition égale des clusters
- Points répartis autour des diamants

=> Méthodologie adaptée

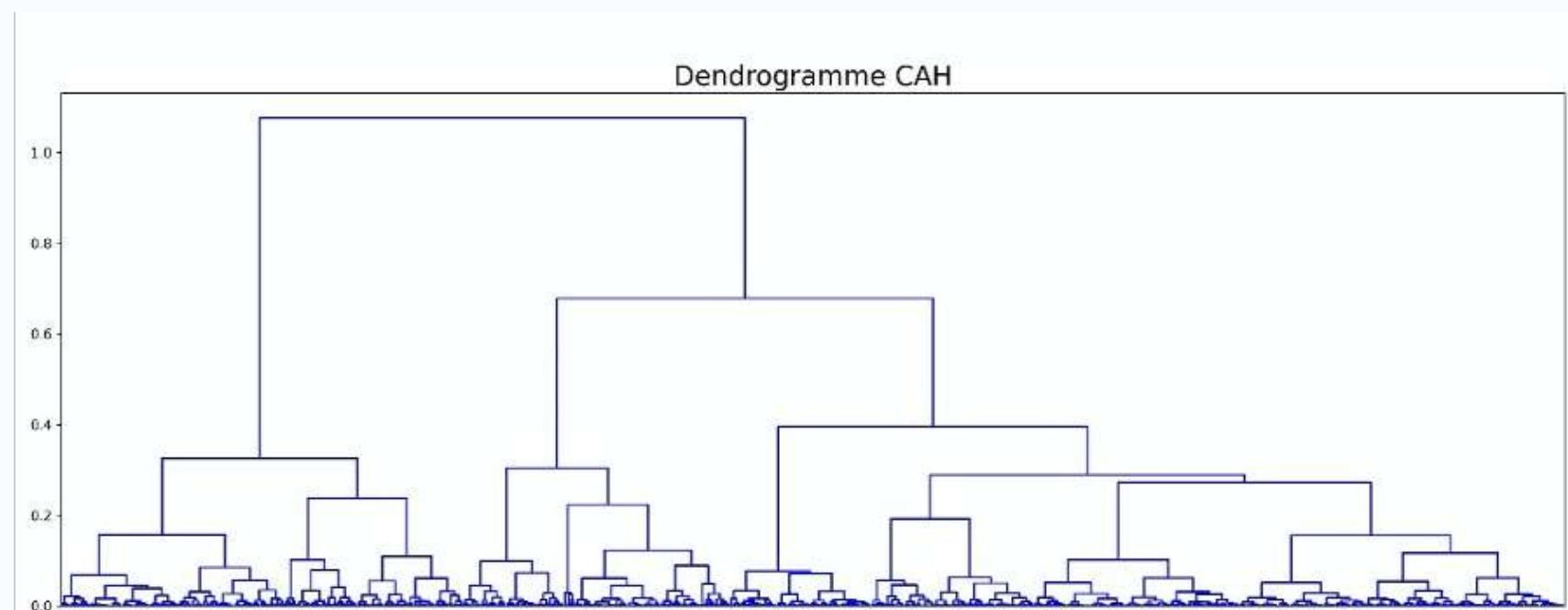


AGGLOMERATIVE CLUSTERING

EN 3 ÉTAPES

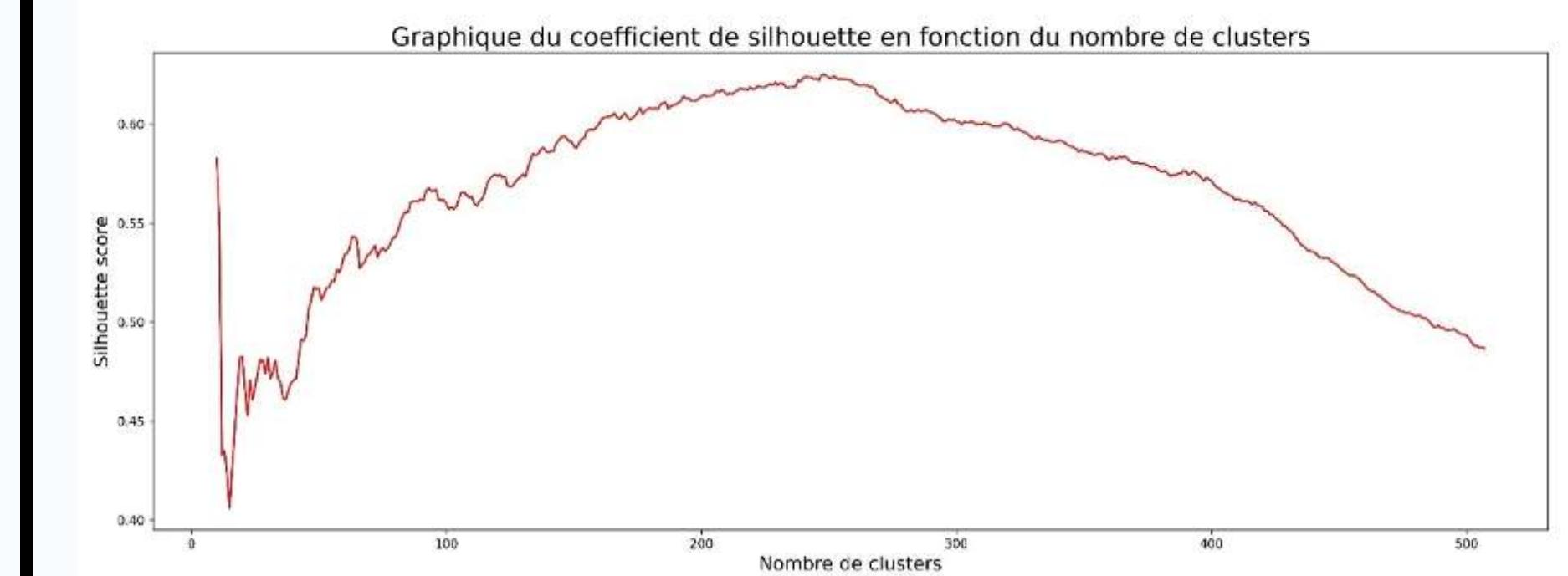


1 - DENDROGRAMME



Pas visible à l'oeil nu

2 - COEFFICIENT DE SILHOUETTE



`n_clusters = 248`

AGGLOMERATIVE CLUSTERING

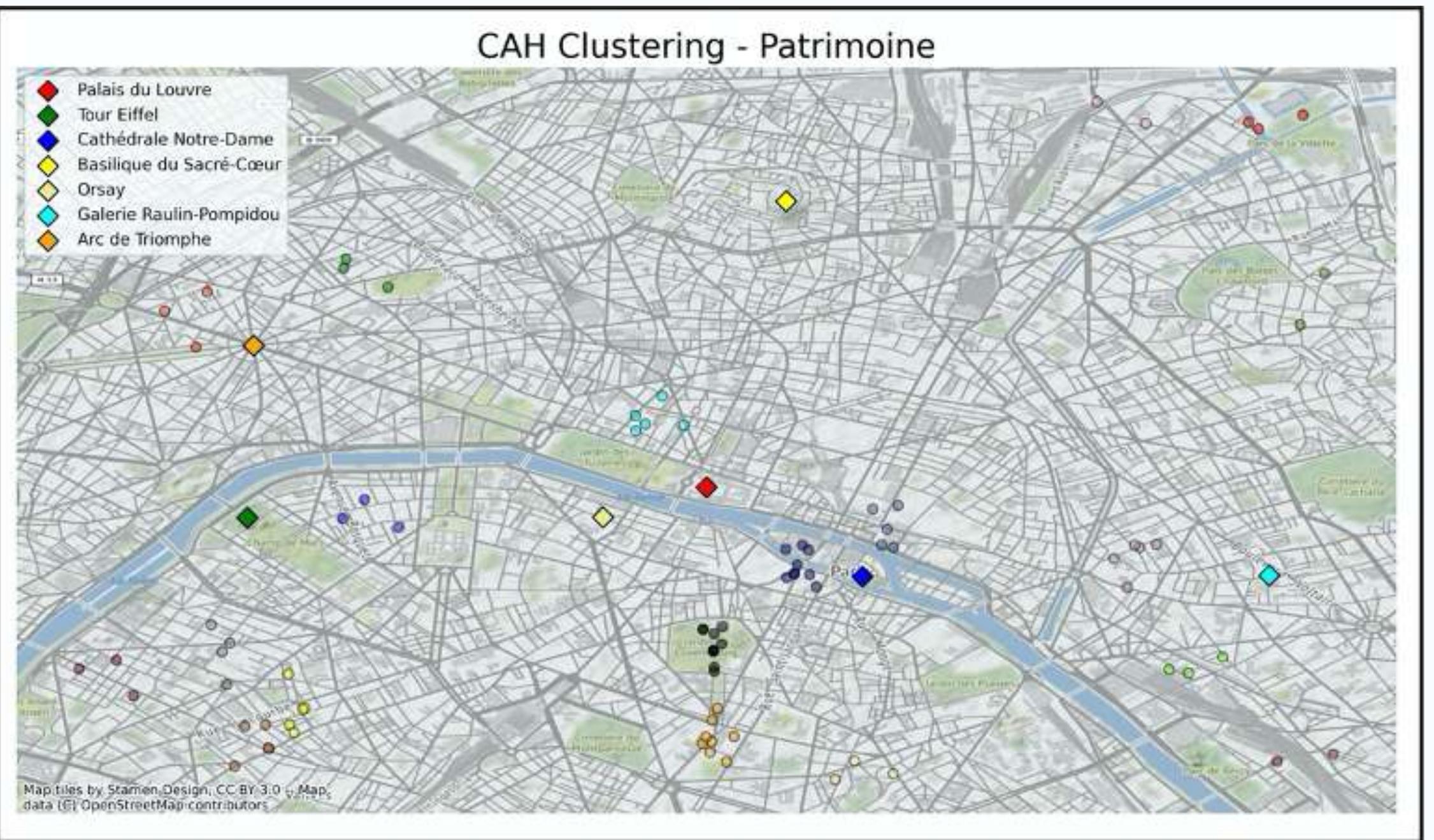
EN 3 ÉTAPES



3 - CLUSTERING ET VISUALISATION

- Beaucoup trop de clusters
- 3 points en moyenne

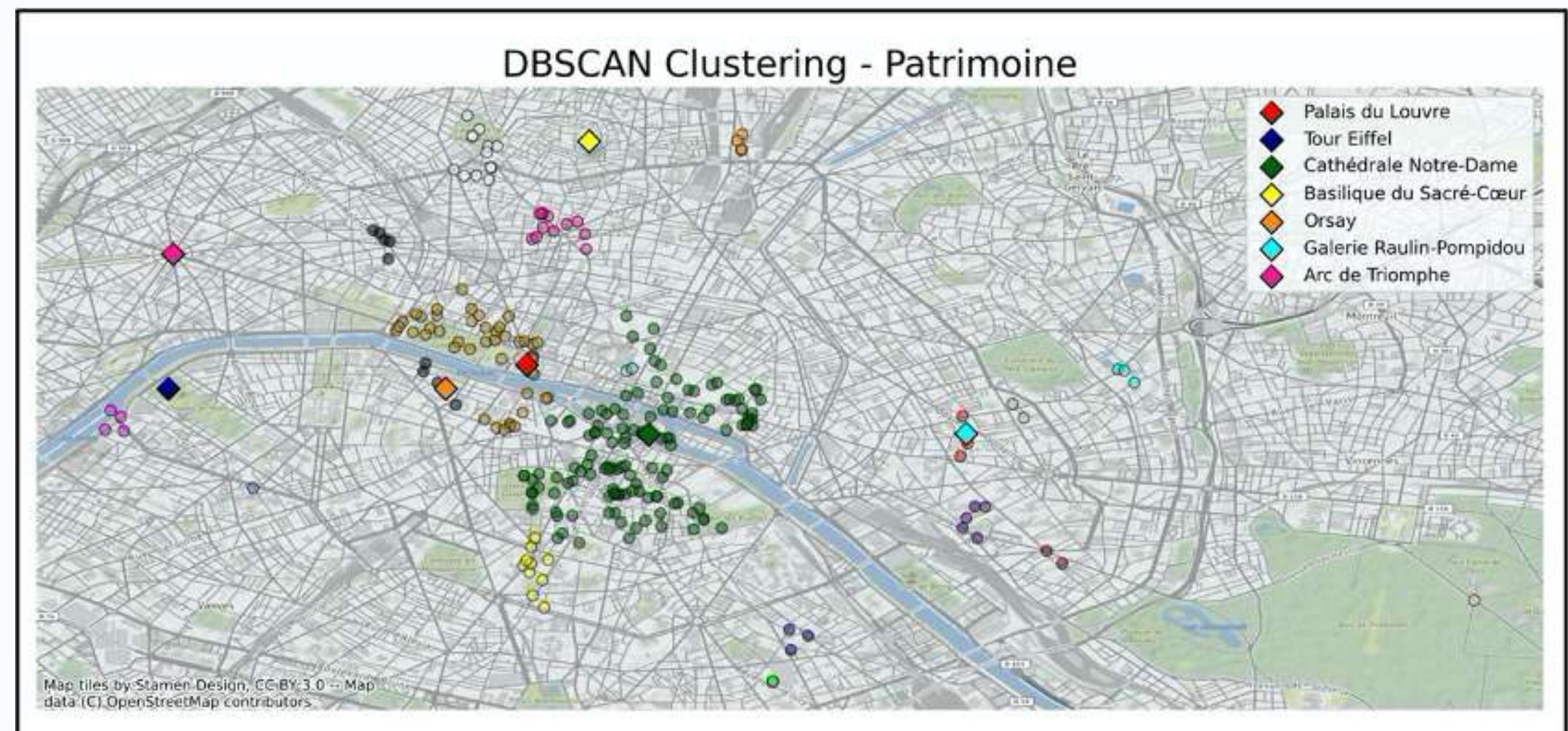
=> Méthodologie peu adaptée au projet



DBSCAN CLUSTERING

DEUX PARAMÈTRES

- Epsilon (*eps*)
 - Nombre minimum d'échantillons pour qu'on considère un cluster (*min_samples*)
-
- Masse de valeurs très élevée
 - Les densités de points sont parfois peu représentatives de la réalité



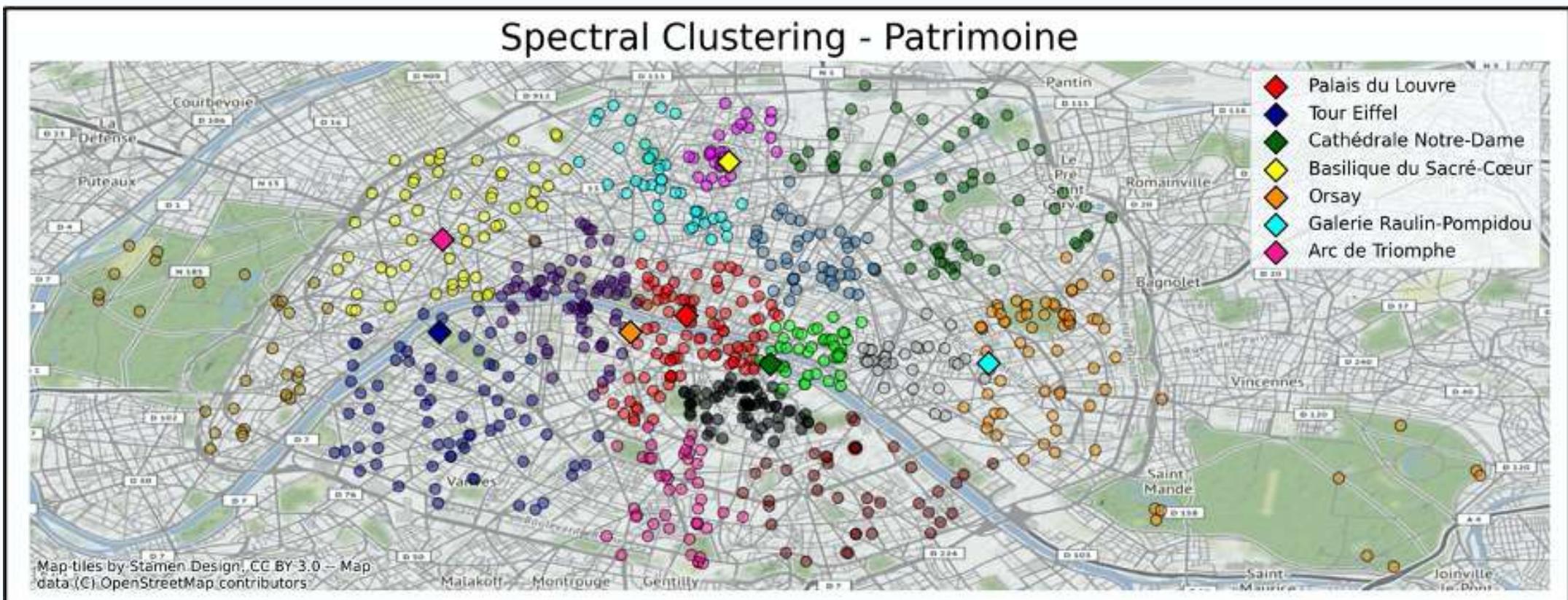
=> Méthodologie peu adaptée au projet

SPECTRAL CLUSTERING

DEUX PARAMÈTRES

- Affinités entre points (*affinity = 'rbf'* ou *'n_neighbors'*)
 - Nombre de clusters (*n_clusters*)
-
- Méthodologie similaire à celle des K-Means
 - Clusters bien formés autour des diamants

=> Méthodologie adaptée



SELECTION DES ALGORITHMES



K-MEANS



AGGLOMERATIVE
CLUSTERING



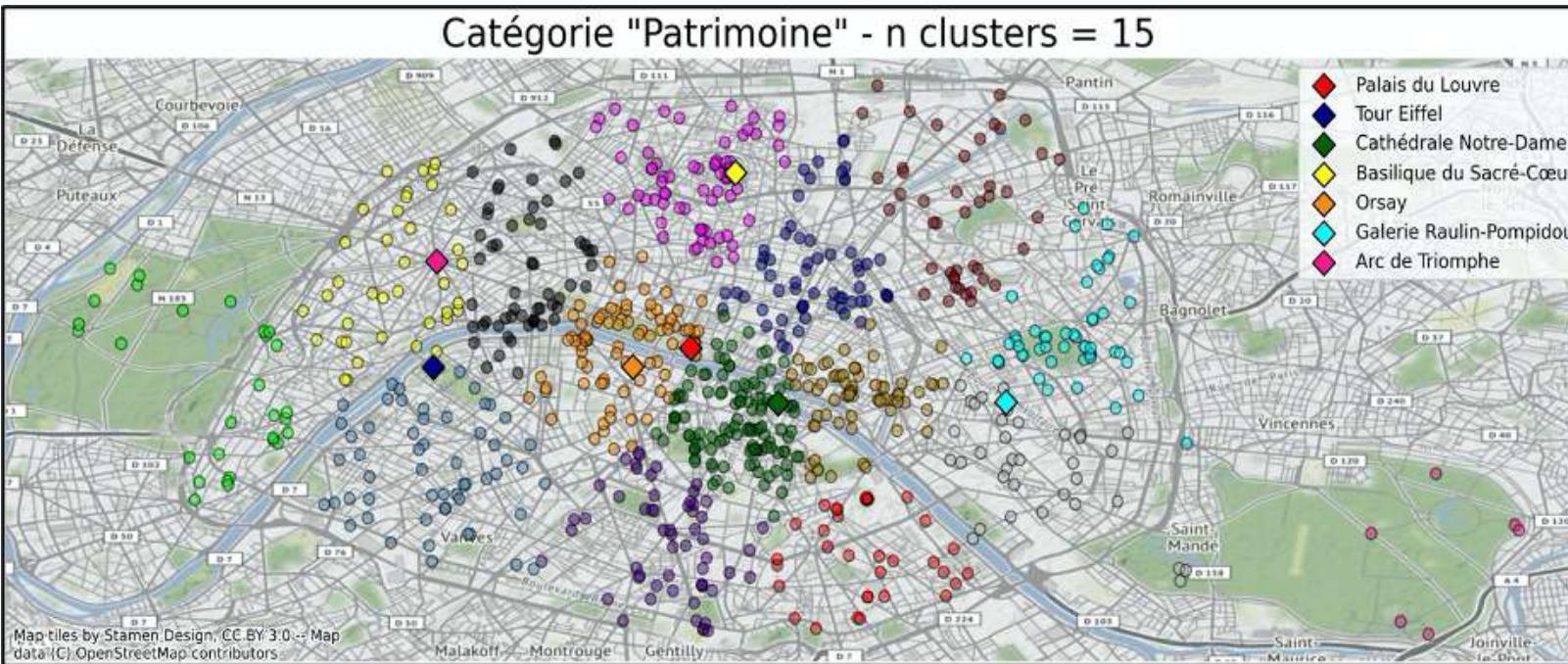
DBSCAN



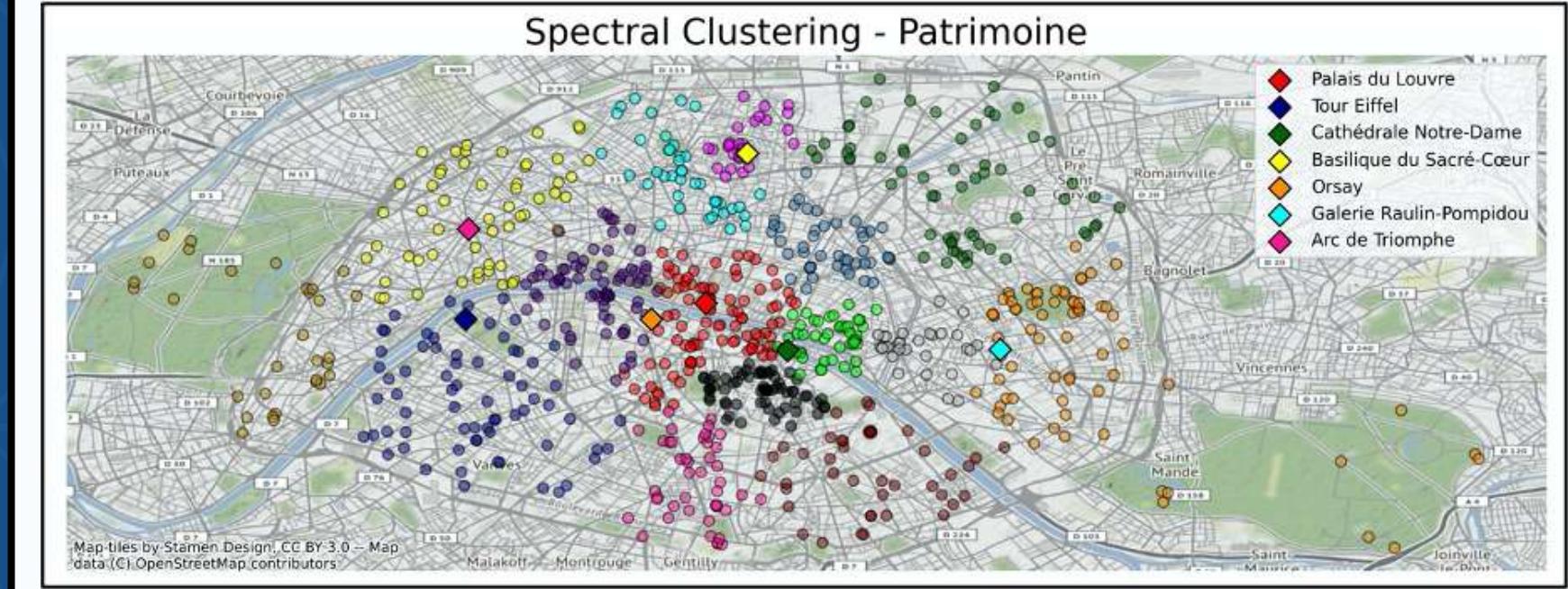
SPECTRAL
CLUSTERING



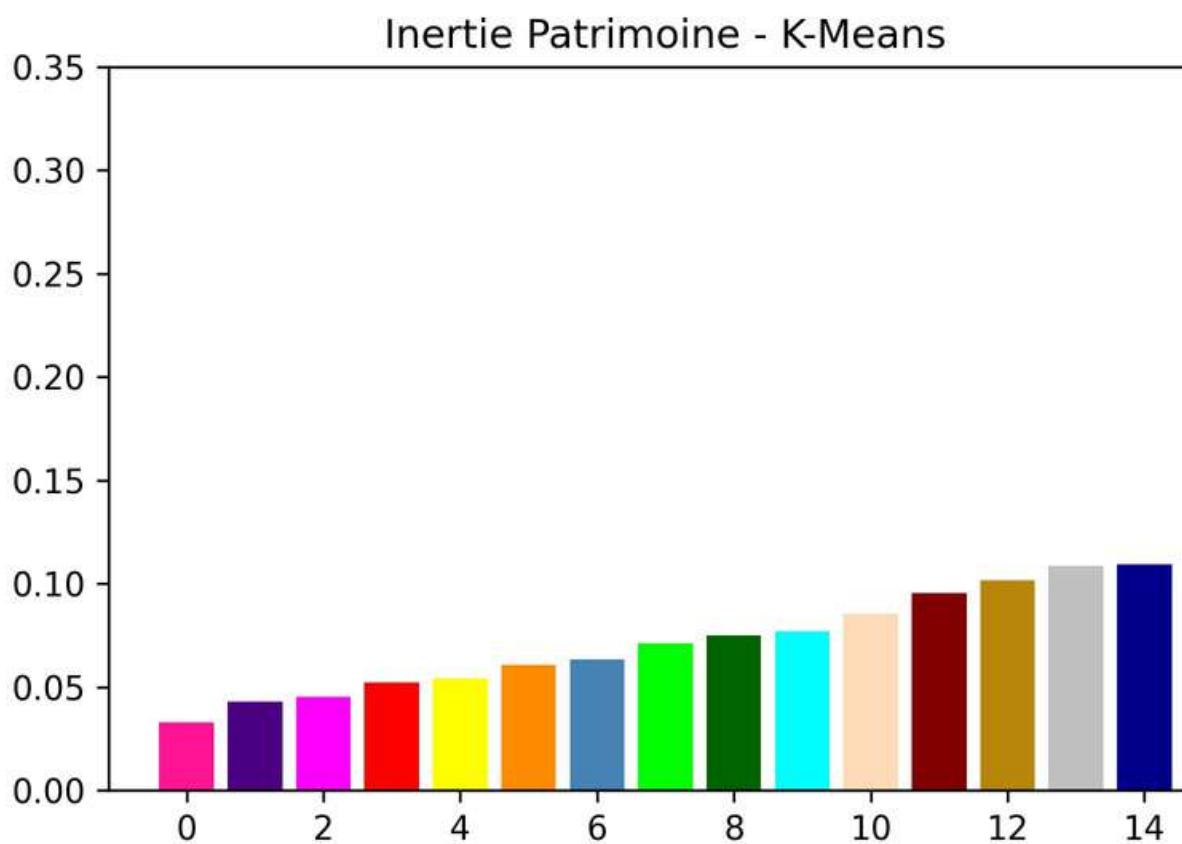
K-MEANS



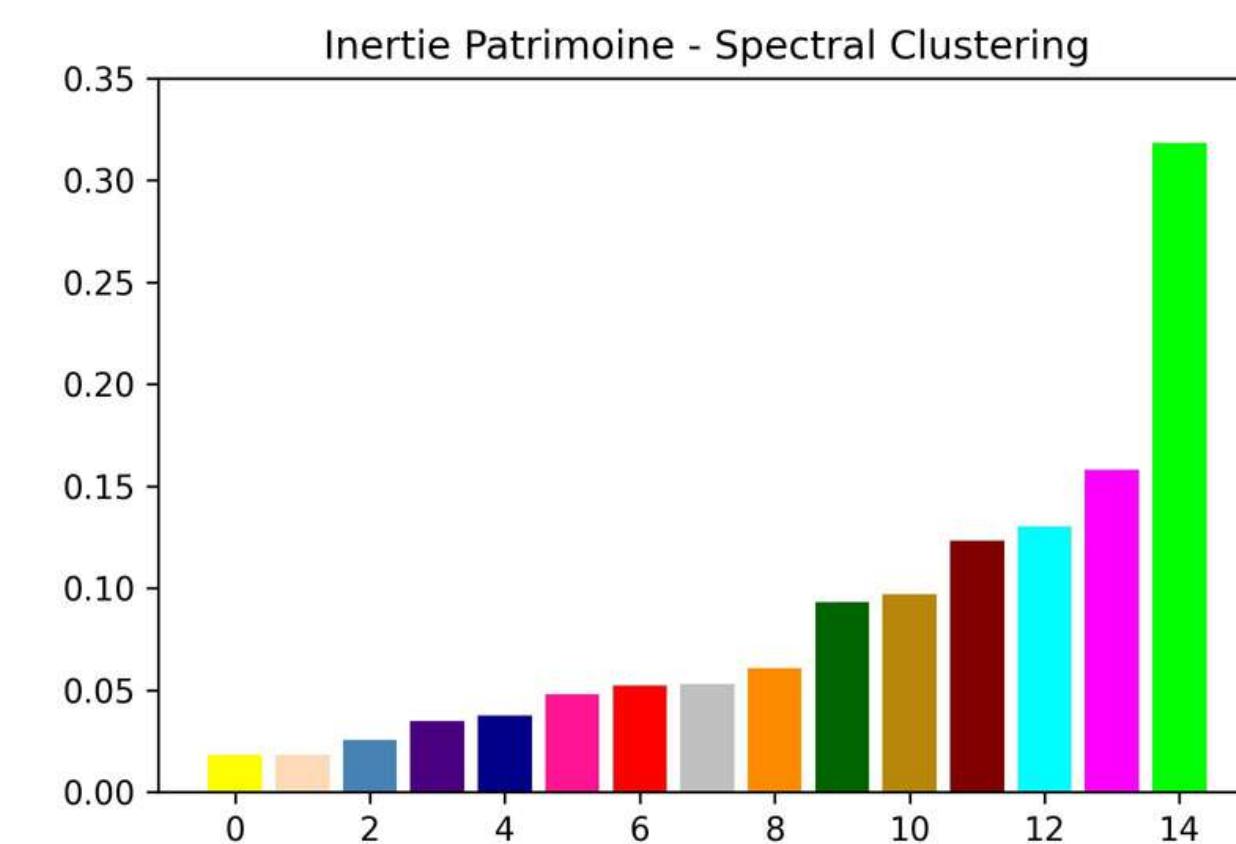
SPECTRAL CLUSTERING



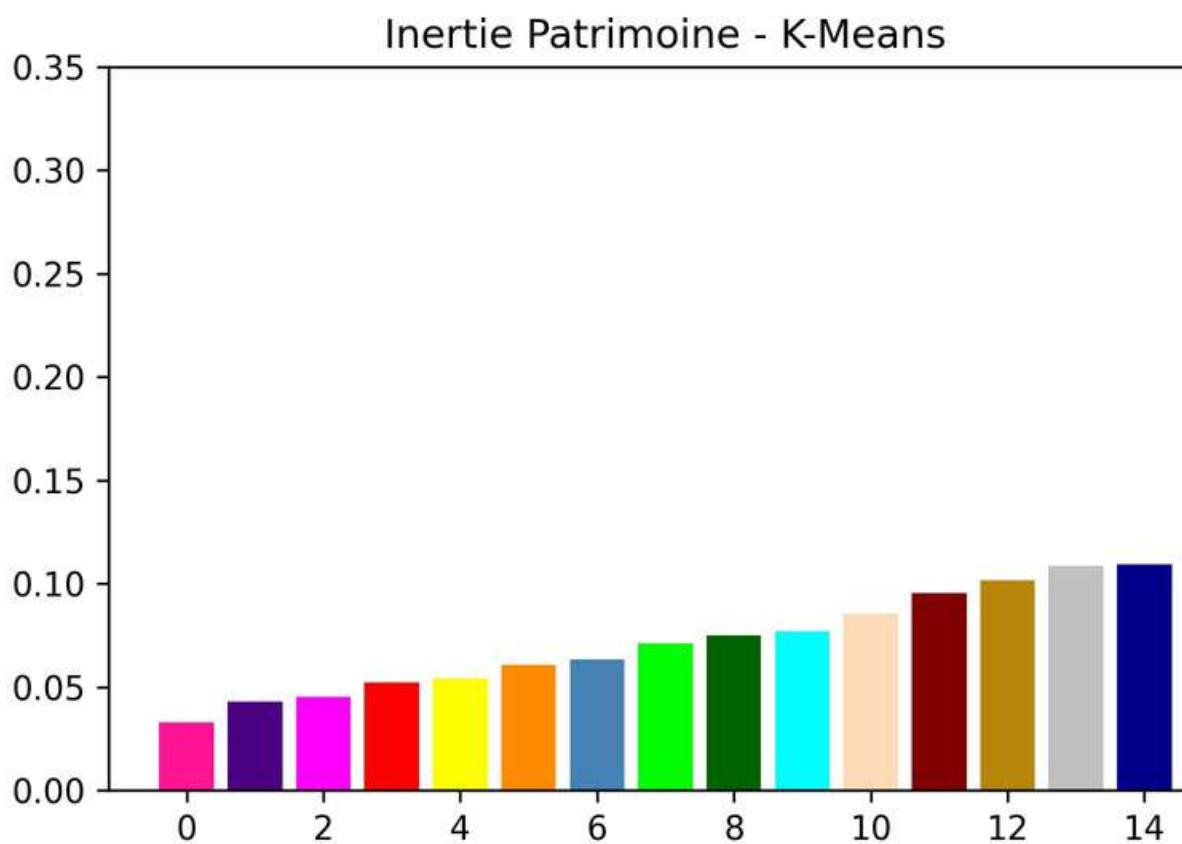
K-MEANS



SPECTRAL CLUSTERING

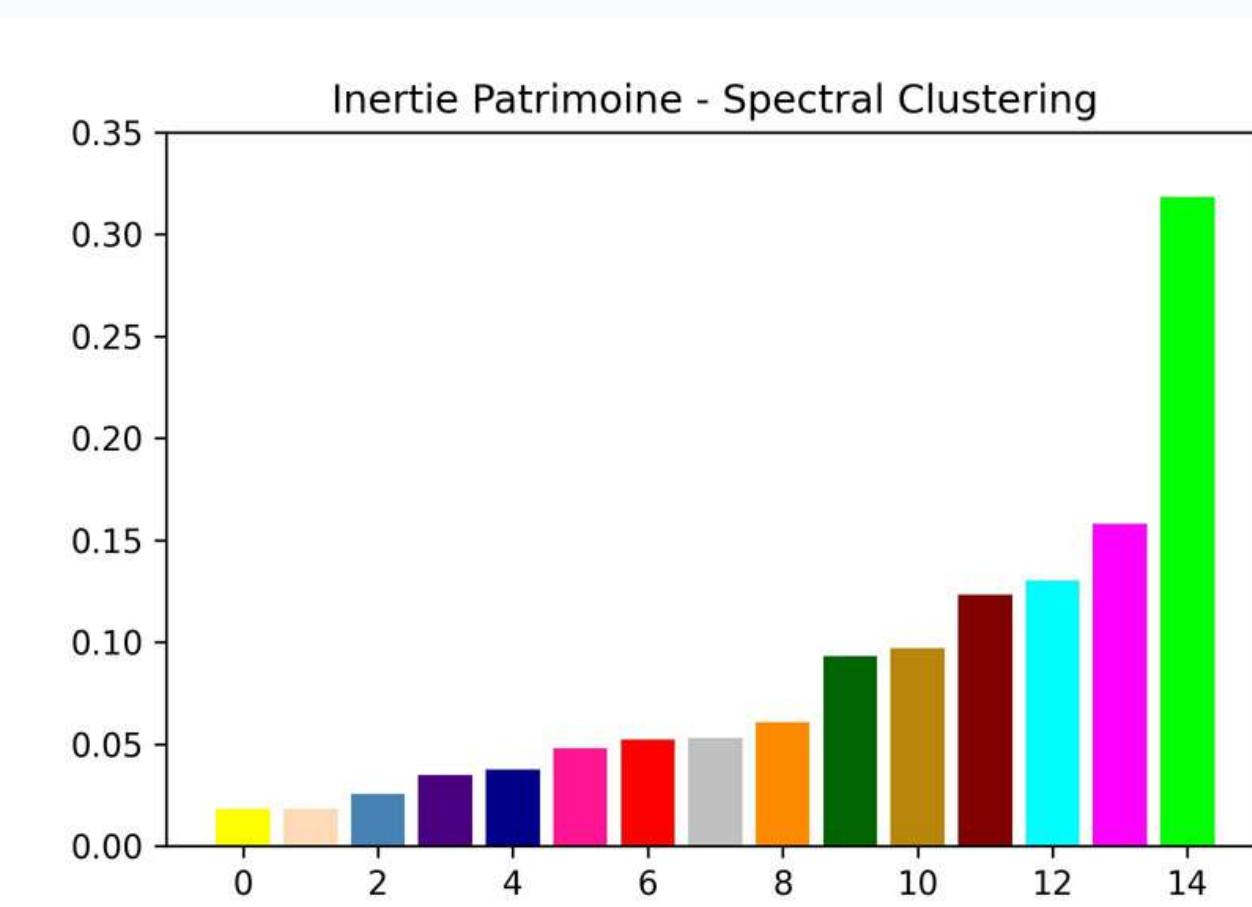


K-MEANS



Pour n_clusters compris entre 10 et 14

SPECTRAL CLUSTERING

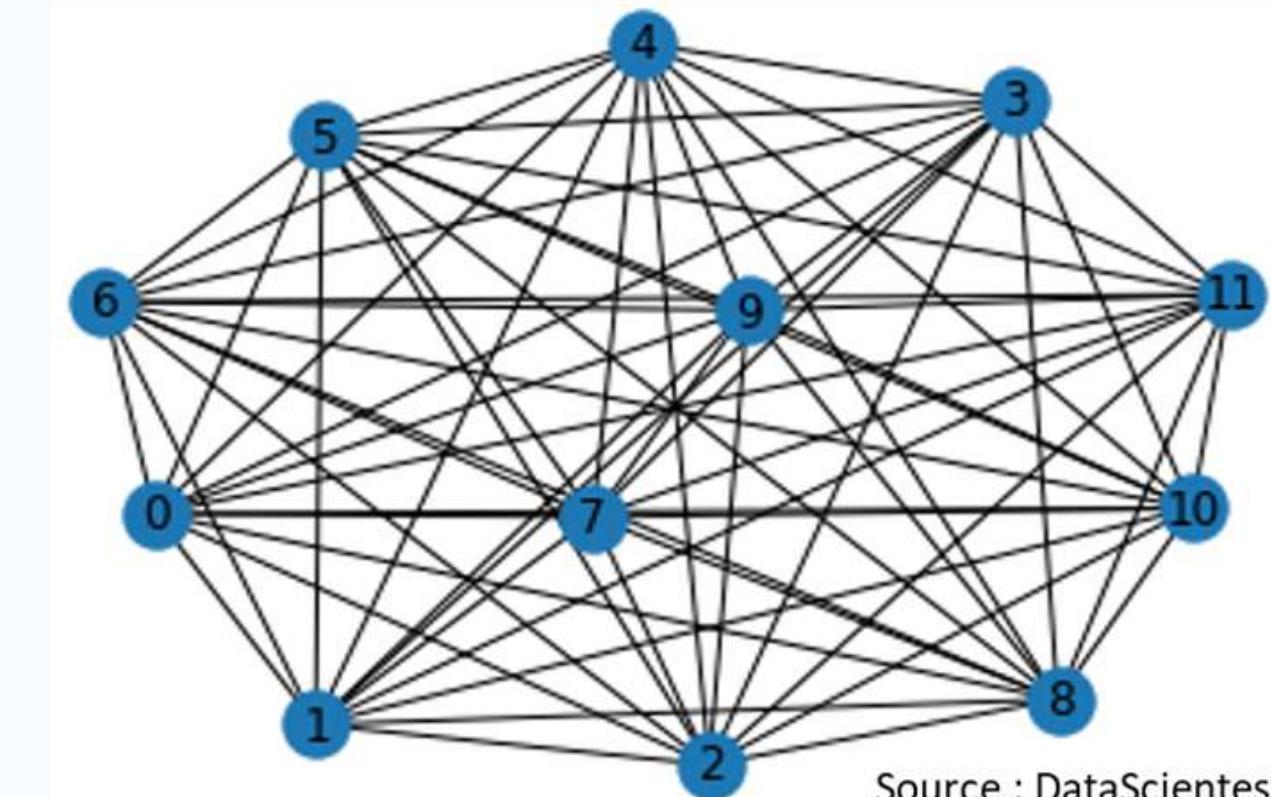


Pour n_clusters compris entre 0 et 9

OPTIMISATION DE L'ITINÉRAIRE : LA THÉORIE DES GRAPHES



- L'étude d'un ensemble de points comme des nœuds (ou nodes) constituant un réseau.
- Ces nœuds peuvent être liés entre eux. Cela est modélisé par des arêtes qui peuvent être pondérées selon divers critères.
- PageRank permet d'identifier les principaux points d'un ensemble et décalculer la probabilité de tomber sur un nœud lorsqu'on circule dans le graphe de manière répétée.



APPLICATION DE PAGERANK À NOS DONNÉES



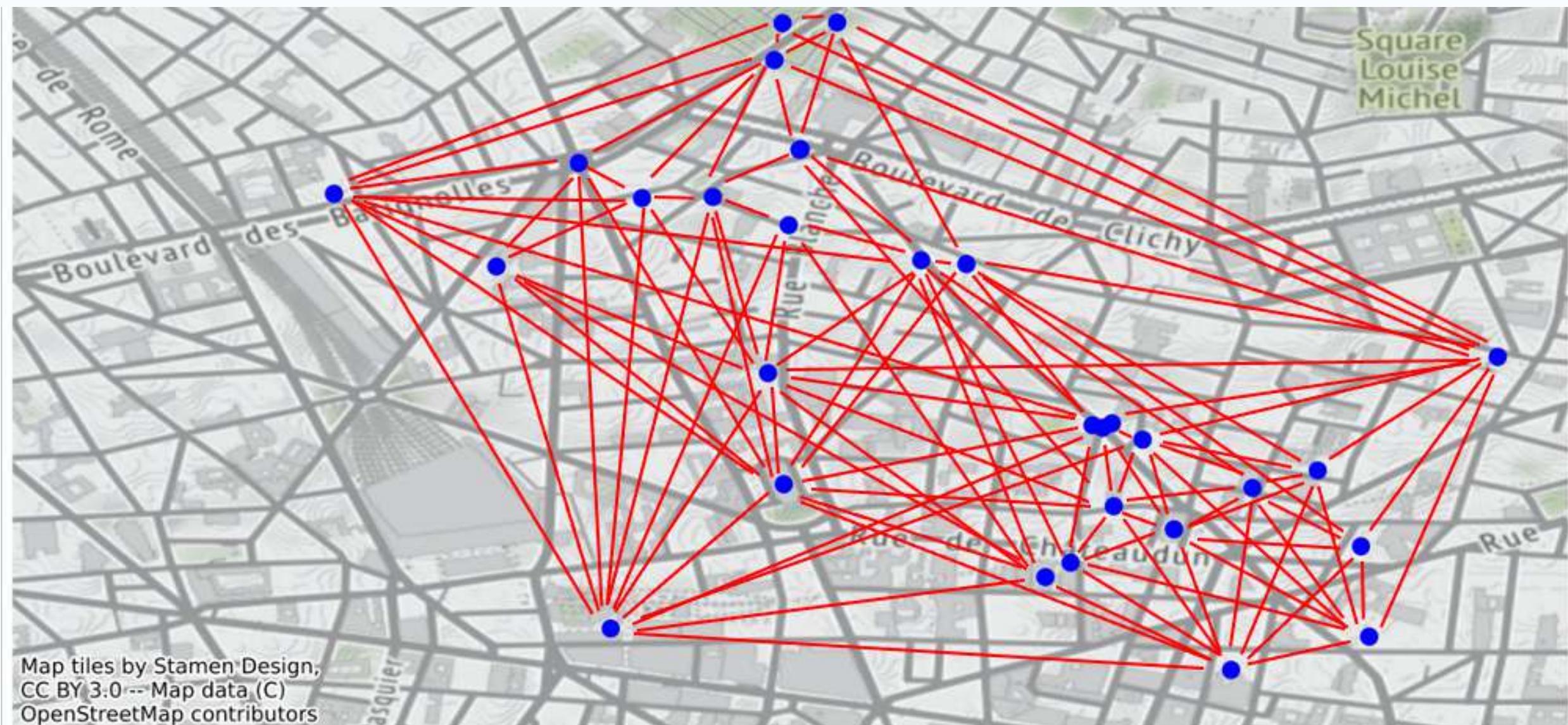
- Notre dataset est composé de points géographiques qui correspondent à des sites patrimoniaux.
- Chaque site est donc considéré comme un nœud d'un même ensemble qui est le cluster défini par le modèle de classification appliqué



APPLICATION DE PAGERANK À NOS DONNÉES



- Le lien entre les nœuds sera la distance qui les sépare.
- Par définition, tous les nœuds sont considérés comme étant liés entre eux.
- Plus la distance entre deux points est réduite, plus la relation sera considérée comme forte.



APPLICATION DE PAGERANK À NOS DONNÉES



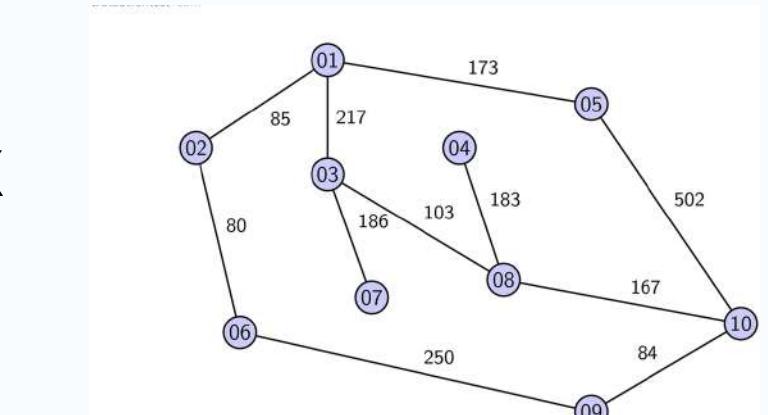
- L'application de PageRank à nos ensembles permet d'identifier les lieux avec la plus grande probabilité d'être visités lorsqu'on circule de manière répétée.
- Il permet de sélectionner des points prioritaires qui optimisent le circuit du touriste.



Tracé des trajets optimaux



- Utilisation de la fonction Minimum Spanning Tree de NetworkX
- Les relations ici sont les distances euclidiennes



Source : DataScientest



FORCES ET LIMITES DE PAGERANK



Il permet une classification par ordre d'importance rapide est cohérente par rapport à des pondérations définies au préalable



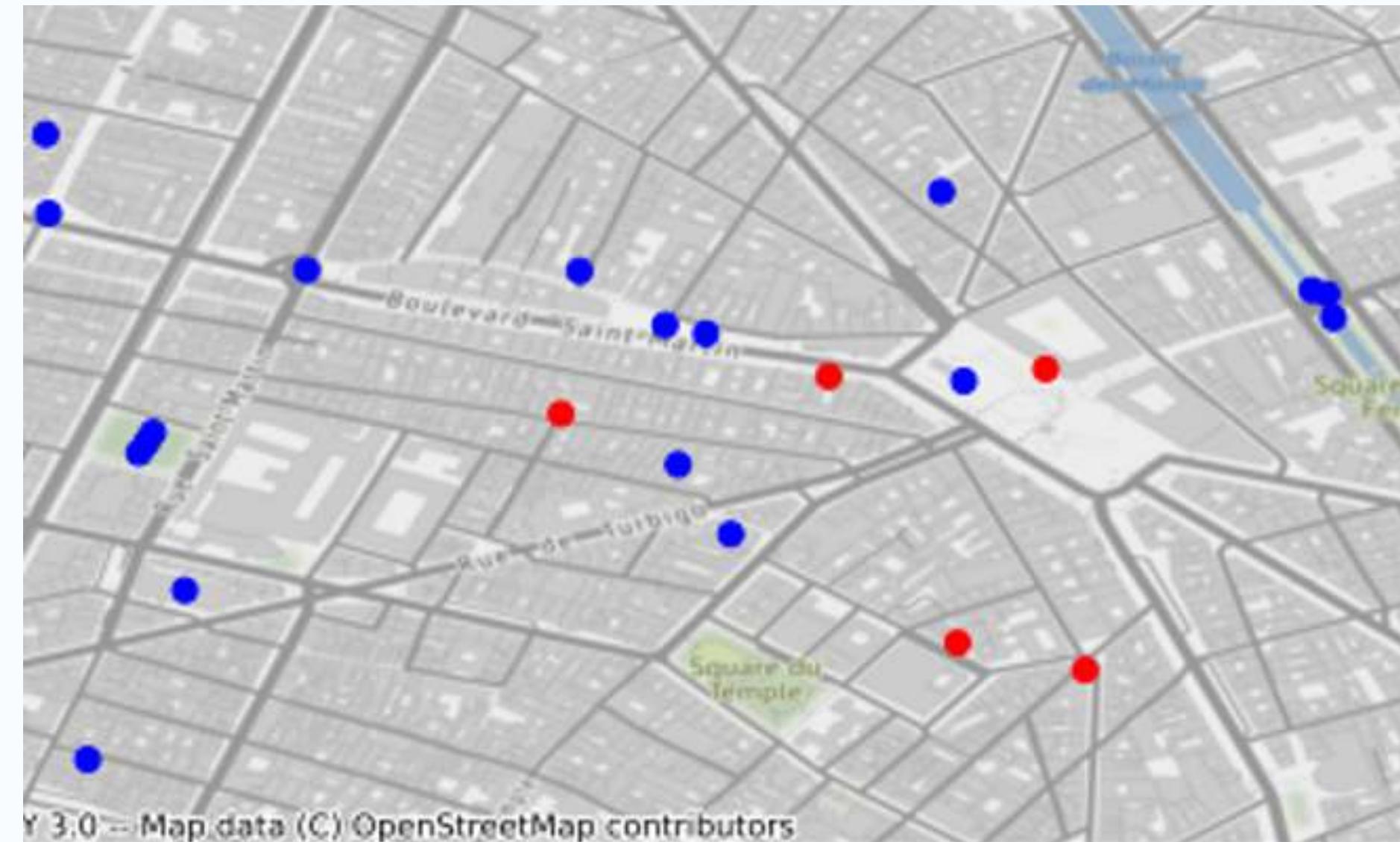
La limite se trouve au niveau du dataset : il manque des informations relatives à l'attractivité des lieux.



INTÉGRATION DES LIEUX SECONDAIRES: RESTAURANTS



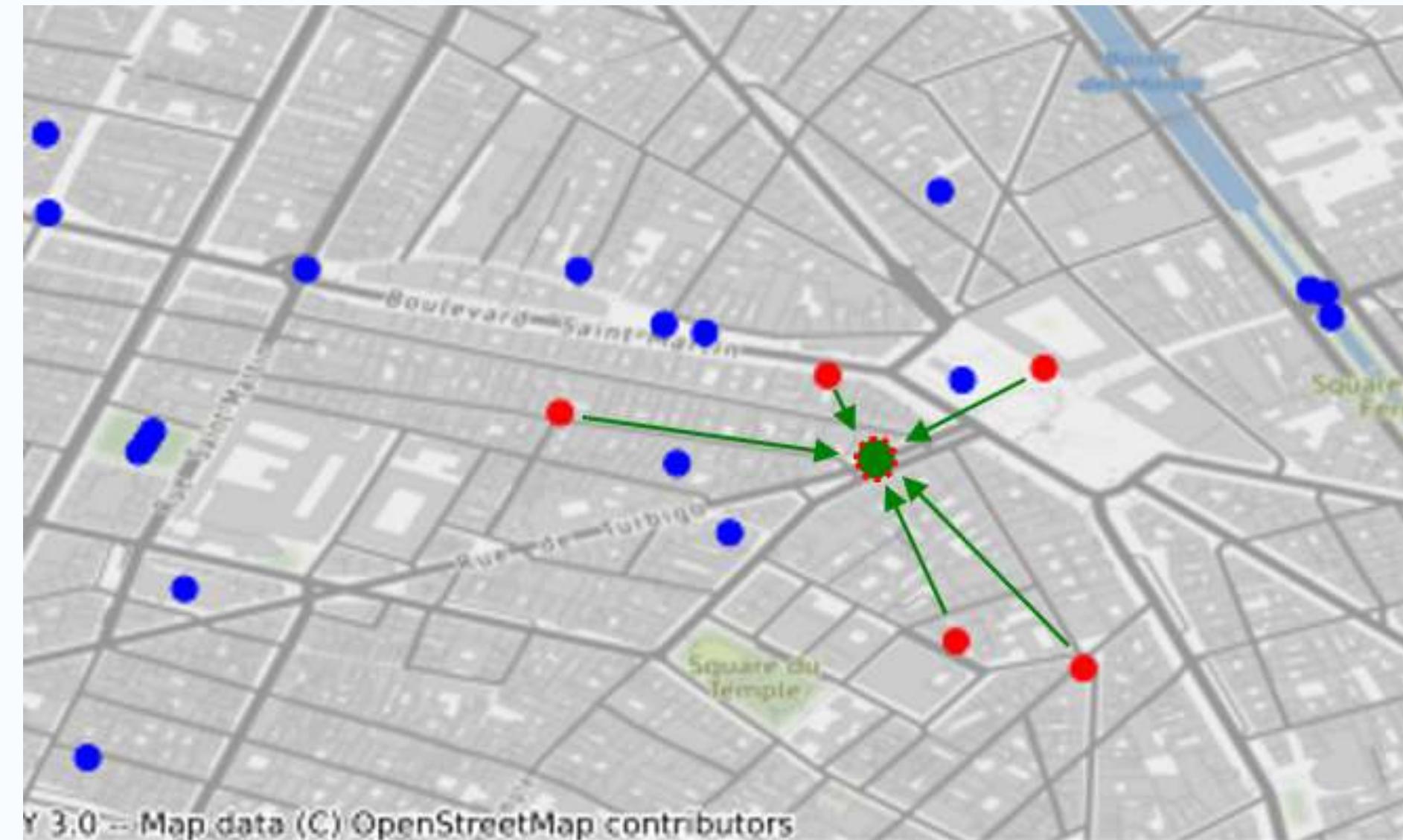
- Les points définis par le PageRank restent prioritaires.
- La sélection des points restaurants dépend de leur proximité au centroïde des points prioritaires.



INTÉGRATION DES LIEUX SECONDAIRES: RESTAURANTS



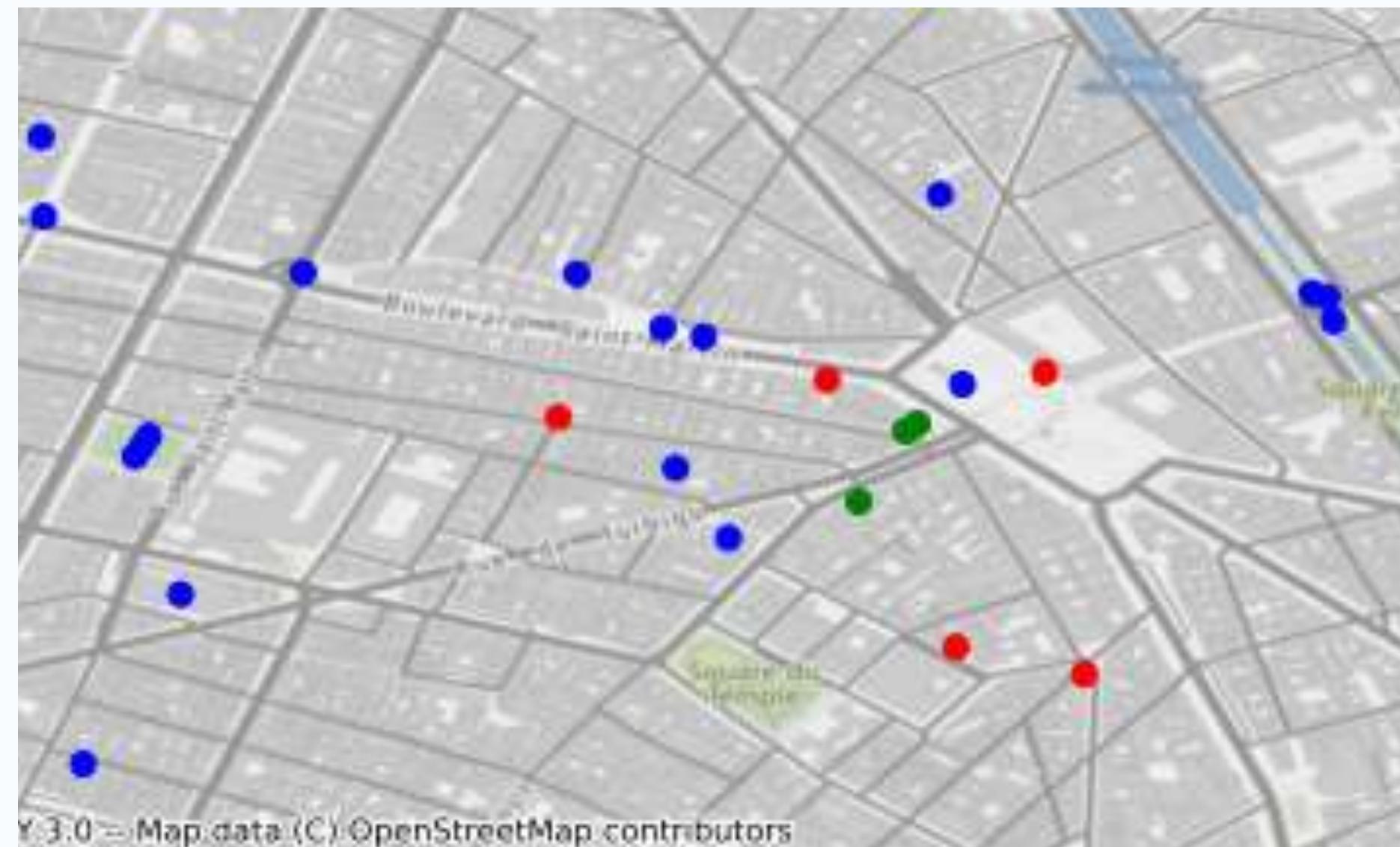
- Les points définis par le PageRank restent prioritaires.
- La sélection des points restaurants dépend de leur proximité au centroïde des points prioritaires.
- Nous calculons le centroïde théorique des points prioritaires



INTÉGRATION DES LIEUX SECONDAIRES: RESTAURANTS



- Les points définis par le PageRank restent prioritaires.
- La sélection des points restaurants dépend de leur proximité au centroïde des points prioritaires.
- Nous calculons le centroïde théorique des points prioritaires
- Grâce à la fonction Cdist du module SciPy, nous affichons les restaurants les plus proches au centroïde



INTÉGRATION DES LIEUX SECONDAIRES: ZONES COMMERCIALES



- L'objectif ici n'est pas d'attirer l'utilisateur vers des boutiques précises, mais vers des zones commerciales.
- Pour identifier des regroupements de commerces, nous avons utilisé un classificateur non supervisé : Optics.
- Pour les restaurants, nous utilisons le centroïde des points prioritaires pour repérer les clusters de commerces les plus proches.



INTÉGRATION DES LIEUX SECONDAIRES: ZONES COMMERCIALES



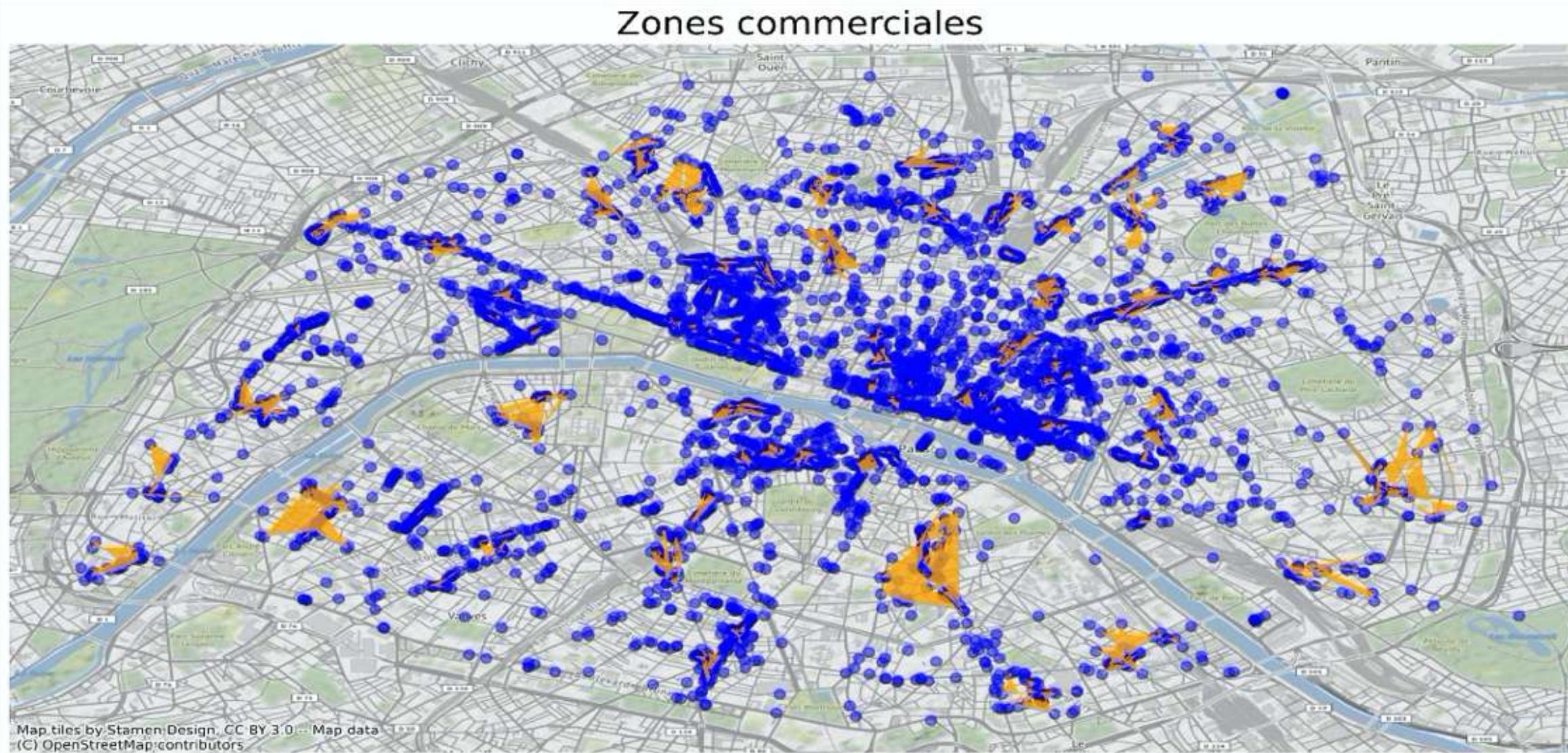
- Pour les restaurants, nous utilisons le centroïde des points prioritaires pour repérer les clusters de commerces les plus proches.
- A l'aide du module Shapely, nous transformons les points en polygones de manière à attirer l'utilisateur vers des zones commerciales plutôt que des points précis.



INTÉGRATION DES LIEUX SECONDAIRES: ZONES COMMERCIALES



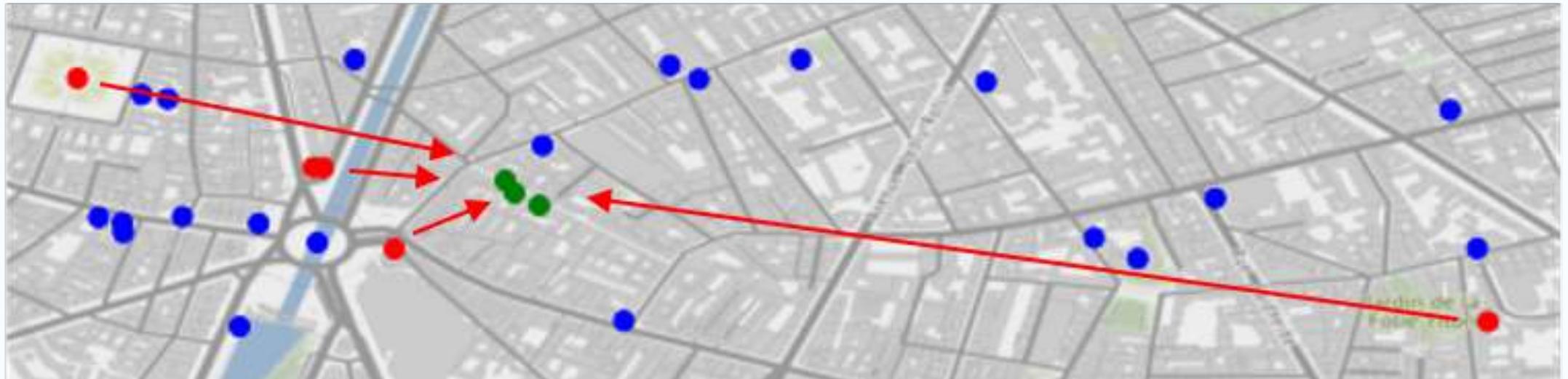
- La performance du modèle est bonne lorsque les zones commerciales sont isolées.
- Cela est moins valable dans des zones commerciales trop vastes



FORCES ET LIMITES DE CETTE DÉMARCHE

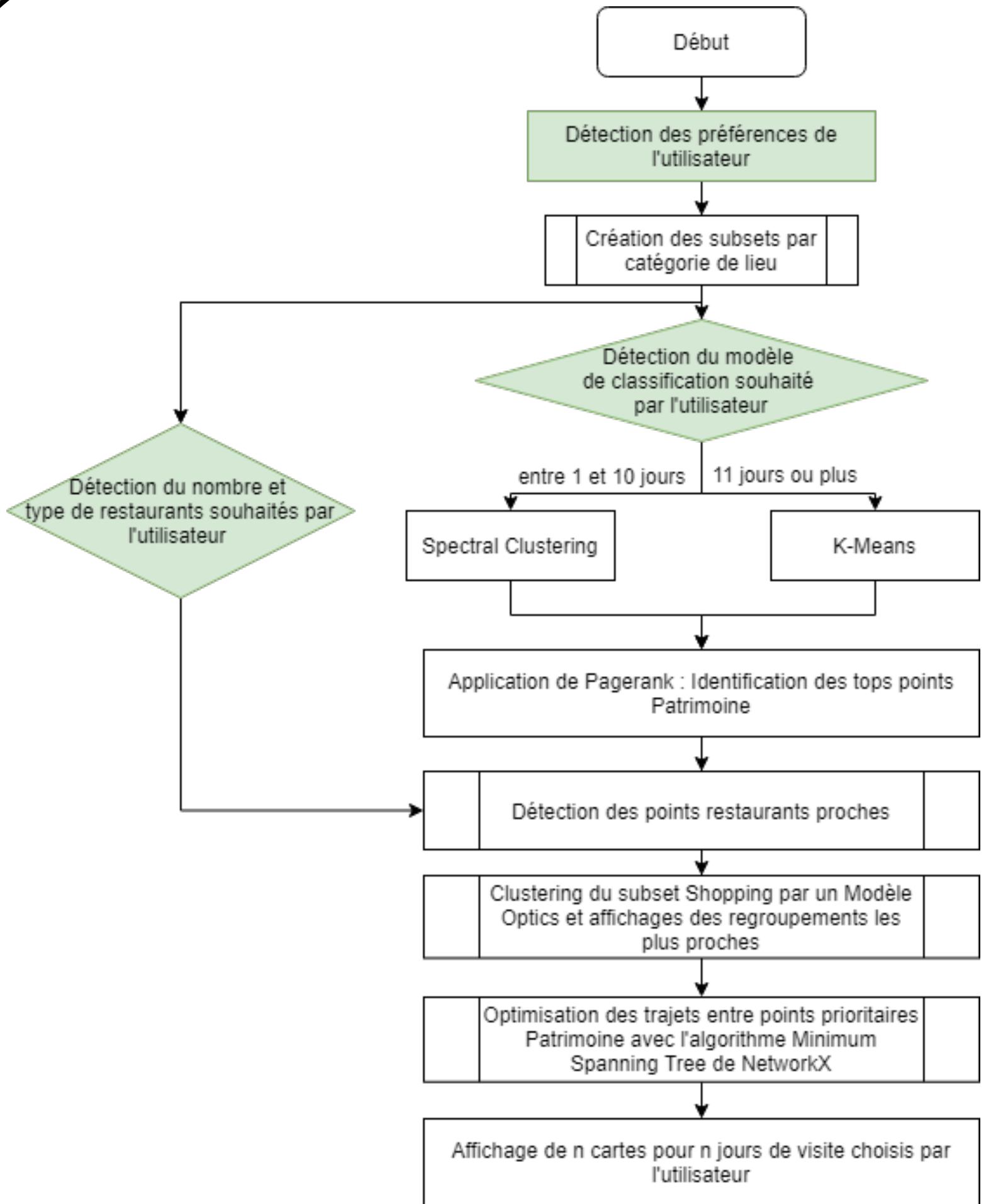


- ▶ Le calcul du centroïde étant sensible aux valeurs extrêmes, lorsqu'un des points prioritaires est excentré, la proposition des restaurants n'est pas optimale.



- ▶ La classification par le modèle Optics a l'avantage permet d'identifier correctement les zones commerciales isolées. Mais le modèle s'avère peu performant lors de zones vastes

ÉTAPES DU STREAMLIT



Merci pour votre attention !



Présenté par Diego Guzman & Danyl Delaisser